FRONTIER TECHNOLOGY INSTITUTE

DATA SCIENCE CERTIFICATION

CAPSTONE PROJECT

Project Title:

Customer Churn Analysis using Machine Learning and Big Data

Project Advisor:

Dr.Tariq Mahmood

Project Manager:

Sania Zafar

Group Members:

Maad Saifuddin

Muhammad Amir

Hussain Murtaza

Syed Ahmed Ali

Danial Khan

Table of Contents:

Introduction:

Objective:

Our main objective is to use machine learning and deep learning techniques to build a model to predict Customer Churn specifically for the Telecommunication Sector.

Customer Churn:

Customer churn happens when users close business with a company or service. Also familiar as customer attrition, customer churn is a nagging measure as compared to gaining new customers because it is much less costly to retain current customers than it is to buying customers, earning business from new customers means working leads the sales funnel and it can cost about 400% to gain new customers. Customer retention is generally more profitable as you've already drawn the trust and loyalty of customers.

Customer churn cutoff growth, so companies will have an interpreted arrangement for managing customer churn in a given duration. By-being conscious and keeping an eye on churn rate, organizations are rigged to regulate their customer retention profit margin and pinpoint policies for advancement. [1]

Procedure to reduce Customer Churn:

There are many possible ways to remove customer churn some are as follows:

1. Concentrate on your loyal customers.

Comparatively focusing on offering incentives to customers who are seeing churning, it could be even more constructive to gather your resources into your trustworthy, useful customers.

2. Investigate churn as it exists.

Analyze your churned customers as a means of concluding why customers are unsubscribing or stop using your services. Analyze when and why churn occurs in a customer's lifespan with your organization, and make that data valuable to put into place appropriate volumes.

3. Pretend that you care for your customers.

Rather than waiting to associate with your customers until they collaborate with you, seek a forethoughtful way. Contact them with all the dividends you provide and show them you care about them, and they'll be implicit to abide. [2]

Importance of churn in Telecommunication:

Churn rate is an influential reflection in the telecommunication service industry. In several geographical areas, many companies clash for gaining new customers, creating opportunities for deportation to make-over from one industry to another.

With expanding burden from rivalry and government authorization, developing retention rates of valuable customers has developed into a gradually serious matter to the telecommunication industry.

Data Description:

About:

Telecom Churn Dataset is from **Orange S.A.**,formerly **France Télécom S.A.**,is a French multinational telecommunications corporation [4] . The data is about the customers in US of Orange S.A

Description:

The "churn-bigml-80" dataset contains 3333 rows (customers) and 20 columns (features).

The "Churn" column is the target to predict.

Each row represents a customer, each column contains the customer's attributes. The datasets have the following attributes or features:

Name of Feature	**Data Type**	**Description**
States	String	Name of states of US
Area Code	Int	Codes in which all the states are divided
Account Length	Int	Amount last credit in their account
International plan	String	Customer is currently active international plan (YES or NO)

Voicemail plan	String	Customer is currently active voicemail plan (YES or NO)
Number Vmail messages	Int	Number of voicemail messages send by the customer
Total day minutes	Int	Number of minutes customer calls in morning
Total day calls	Int	Number of calls of Customer In the morning
Total day charge	Int	Totals charges of calls of customers in morning
Total eve minutes	Int	Number of minutes customer calls in evening
Total eve calls	Int	Number of calls of Customer In the evening
Total eve charge	Int	Totals charges of calls of customers in evening
Total night minutes	Int	Number of minutes customer calls in night
Total night calls	Int	Number of calls of Customer In the night

Total night charge	Int	Totals charges of calls of customers in night
Total intl minutes	Int	Total Number of international minutes
Total intl calls	Int	Total number of International calls
Total intl charge	Int	Totals charges of International calls
Customer service calls	Int	Number of times calls on customer service center
Churn	String	Customer Churn (Yes or No)

Checking null Values:

There were no missing values in the data.

Statistical Analysis:

Cardinality:

State 51

Account length 212

Area code 3

International plan 2

Voice mail plan 2

Number vmail messages 46

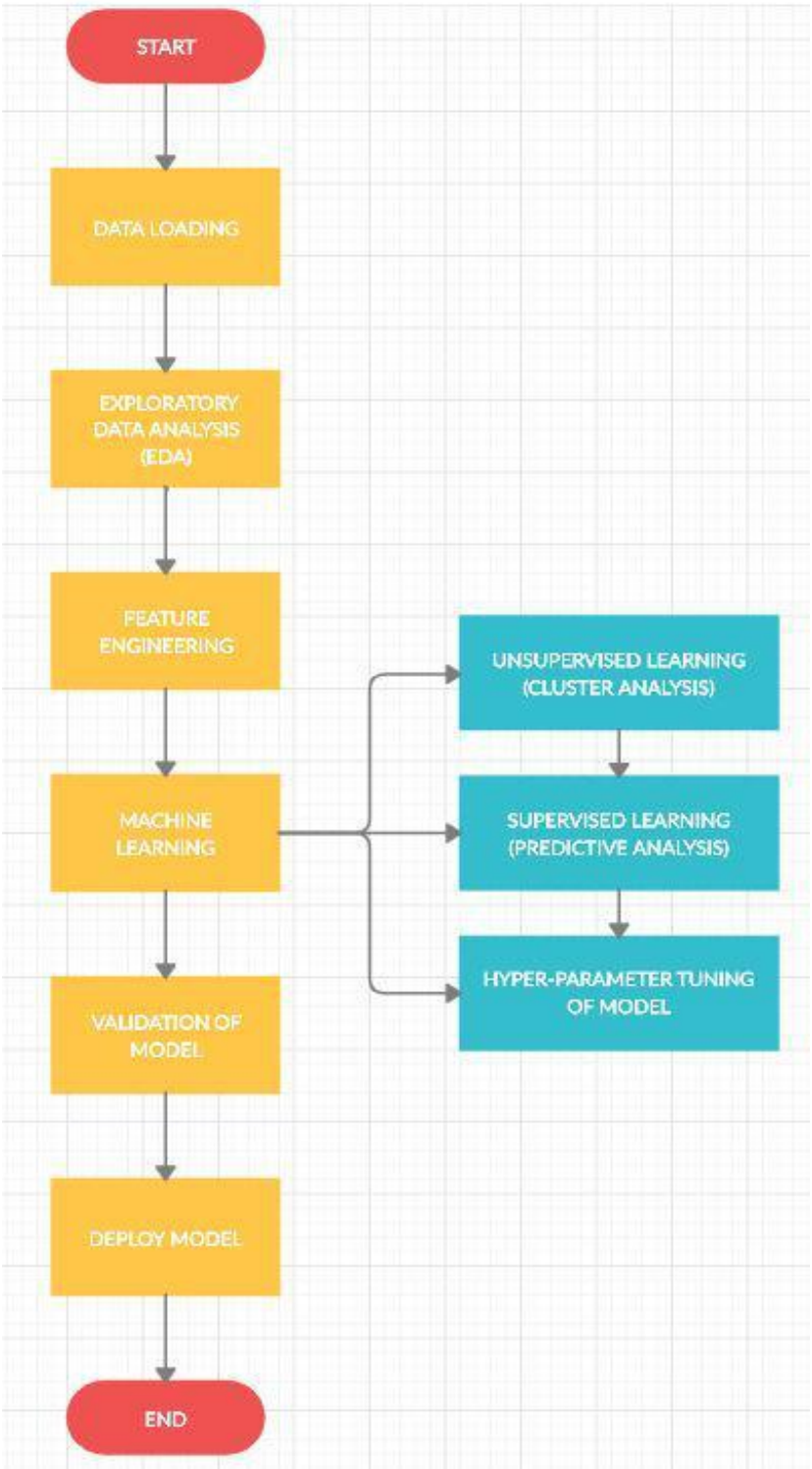
Total day minutes 1667

Total day calls 119

Total day charge	1667
Total eve minutes	1611
Total eve calls	123
Total eve charge	1440
Total night minutes	1591
Total night calls	120
Total night charge	933
Total intl minutes	162
Total intl calls	21
Total intl charge	162
Customer service calls	10
Churn	2

Area code,international plan,Voicemail plan,Total intl calls,Customer service calls has low cardinality While Other Variable has high cardinality.

Flow Chart of Work Flow:



EXPLORATORY DATA ANALYSIS (EDA):

- **UNIVARIATE ANALYSIS:**

The dataset contains 5 categorical attributes ("State", "Area code", "International plan", "Voicemail plan", "Customer service calls") and 14 numerical attributes ("Account length", "Number vmail messages", "Total day minutes", "Total day calls", "Total day charge", "Total eve minutes", "Total eve calls", "Total eve charge", "Total night minutes", "Total night calls", "Total night charge", "Total intl minutes", "Total intl calls", "Total intl charge").

Categorical Attributes:

State: It belongs to 53 states of the US and it contains 53 values out of which maximum users are coming from West Virginia (WV) i.e. 106, Minnesota (MN) is at number second and has 84 users and New York (NY) is at number third with 83 users.

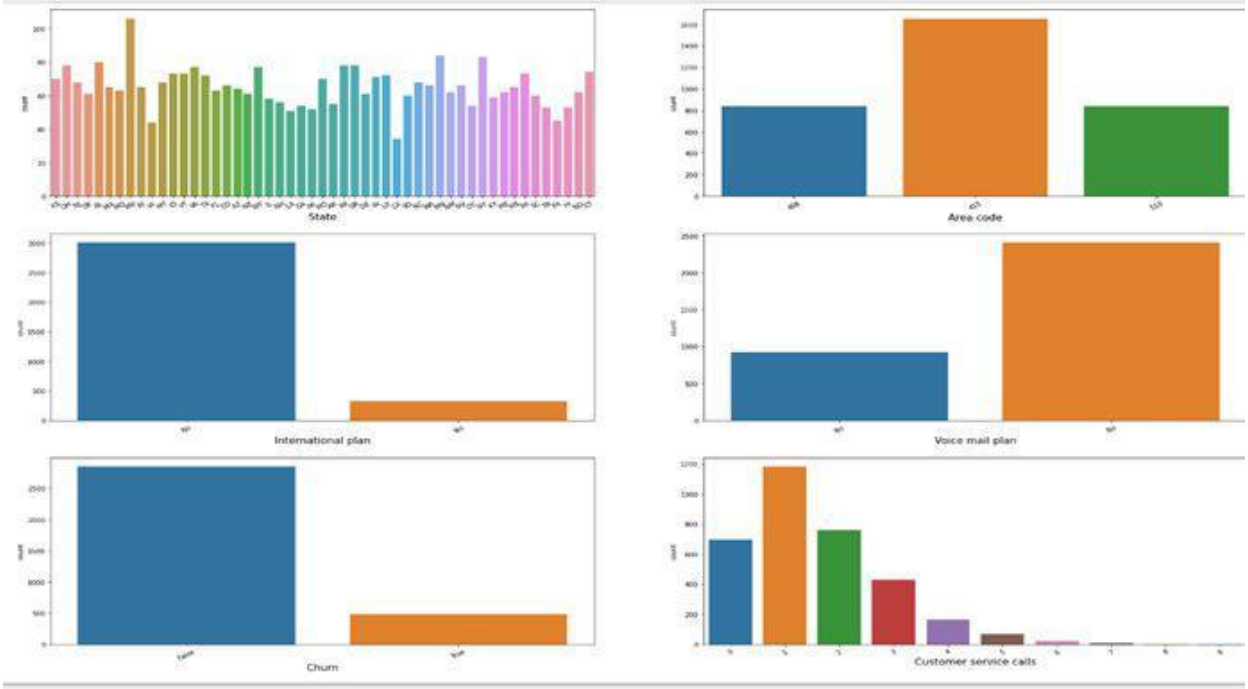
Area Code: The users of US are divided into 3 areas i.e. 415, 510 and 408 out of which maximum users belong to 415 area code with a number of 1655 and second area from which the users are coming is 510 with a number of 840 and third area from which the users are coming is 408 with a number of 838.

International Plan: Most of the users are not using international plans.

Voicemail Plan: Most of the users are also not using voicemail plan.

Churn: There are a smaller number of users that are churning.

Customer Service Call: The highest number of customer service calls received by the user is 1.



Numerical Attributes:

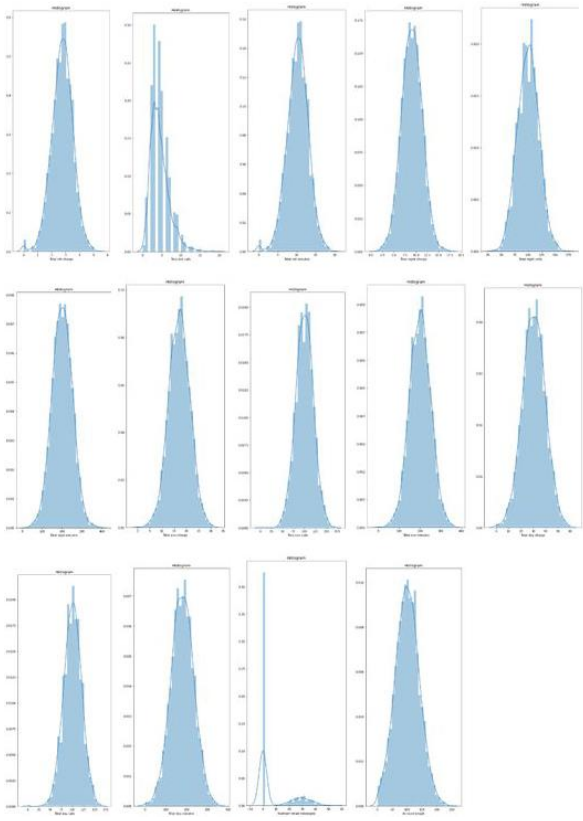
Account length: The user account length means how much users are using their service; the shape of the attribute is in Gaussian form.

Number of voicemail messages: The attribute contains the information that how much users are using their voice mail service, the shape of this attribute is not Gaussian.

Total day minutes, total night minutes, total evening minutes and total international minutes, this attribute contains the information that how much a user uses minutes when he calls someone in their whole day and international minutes represent how much the user uses their service to call someone internationally. The shape of all attributes is Gaussian.

The third form of attribute is call which includes total day calls, evening calls, night calls and international calls. How much the user received a call in the whole day, the form of this day is also Gaussian.

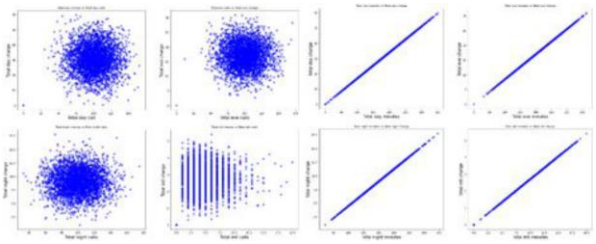
Total day charge, Total evening charge, Total night charge and total international this attribute represent how much users load the charge in their phone.



Bivariate Analysis:

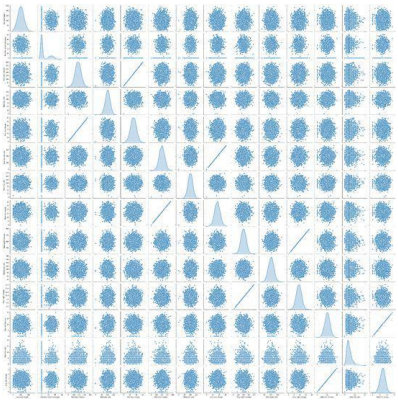
From the Below Graph we conclude that some features are positively correlated with each other and some have no pattern in it means they are neutrally correlated with each other.

The features which form positive correlation are Total day minutes and total day charge, total evening minutes and total evening charge, Total night minutes and total night charge, Total international minutes and total international charge.



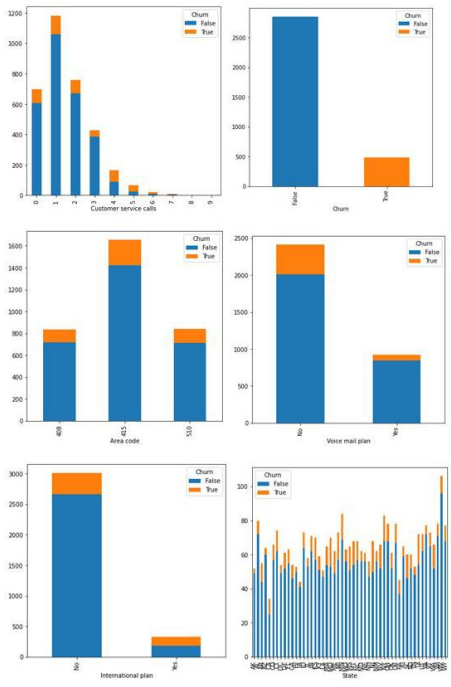
Numerical to Numerical Relation:

The Multivariable Numerical features graph is below which represents every single feature relation with all numerical features.



Categorical to Categorical:

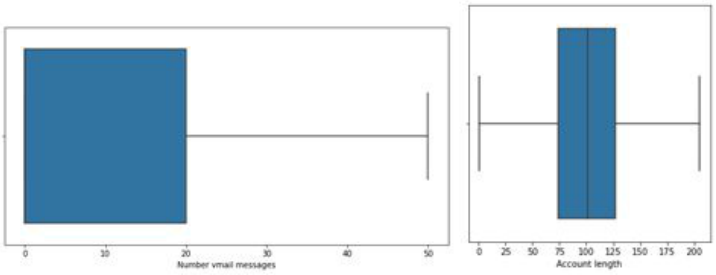
Below Graph represents the relationship between categorical to target. As we see below, all categorical features represent that we have less churners than a churner.



Outlier Engineering:

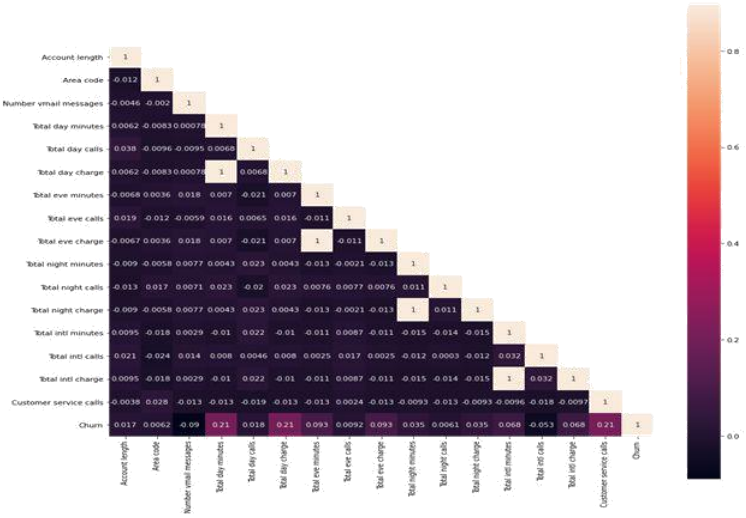
We used z_score interquartile range (IQR) to detect outliers and we found that there were 415 outliers present in the dataset. We have used the technique of replacing outliers with mean instead of removing it.

All outliers were replaced with mean but here we only show 2 graphs after outlier engineering. Some graphs after outlier engineering are below.



Feature Engineering:

We first plot a multicollinearity graph to see the correlation of features with the target.

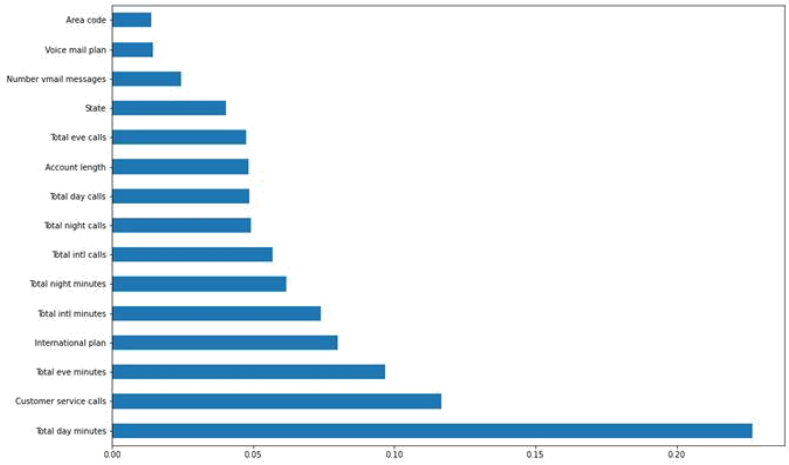


From the above Graph we analyze that some features are highly correlated with each other as we mention above in bivariate analysis, so we remove all types of charge features from our data sets.

From above we analyze that Total day minute, Total day charge and customer service calls are highly positive correlations in range of 0.21 with target churn, whereas Number vmail message and international plan features are negatively correlated with target churn.

First we save our data frame into a new data frame, for Feature Selection we use Wrapper Method Random Forest Classifier to select the feature, which has more importance with target variable then other feature. We use the threshold of less than equal to 0.0250; features which were not selected are Area code, Number of voice mail messages and voicemail plan.

Below Graph represents the importance of features with respect to target.



Feature Selected By Random Forest

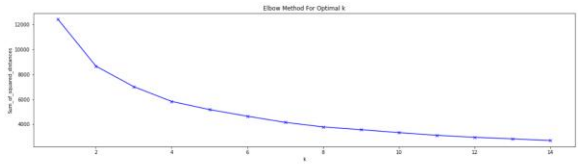
Methodology:

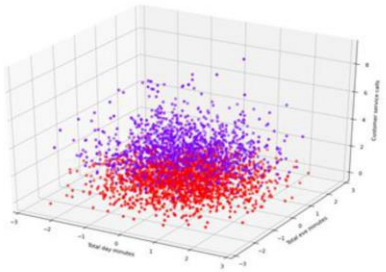
Un-Supervised Learning:

First, we apply clustering and did cluster analyze we apply different technique for clustering like Density based clustering, affinity propagation, hierarchical agglomerative clustering, K mean and mean shift Algorithm. For clustering we take top 3 features which are highly correlated and more important with target churn.

K-Means: In K-means, we use the elbow method to find how many clusters we want to make, there are two clusters formed.

Elbow Method:





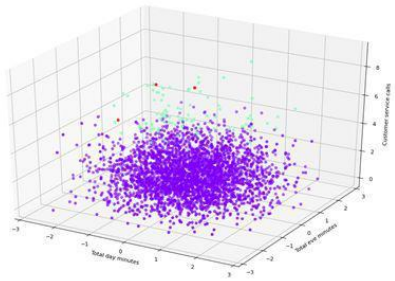
Cluster	Total day minutes	Total eve minutes	Customer service calls
0	-0.0	-0.0	2.8
1	0.0	0.0	0.6

The table shows centroids of each cluster that could determine the clusters rule. These are:

Cluster 0: Negative total day minutes, negative total eve minutes, more customer service calls, and short duration customers.

Cluster 1: Positive total day minutes, positive total eve minutes, less customer service calls, long duration customers.

Mean Shift: In mean-shift, we tuned our bandwidth to get the best parameter for cluster and we analyze that at quantile is equal to 0.2 we found three cluster.



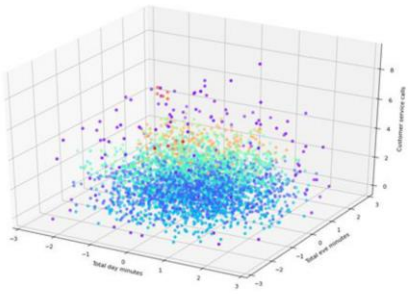
Cluster	Total day minutes	Total eve minutes	Customer service calls
0	0.0	0.0	1.4
1	-0.1	0.1	5.5
2	-0.0	-1.4	7.2

Cluster 0: represent positive zero values they are not churner.

Cluster 1: represent negative value of day minutes and positive zero of eve minutes and little high customer service calls means they are less churner.

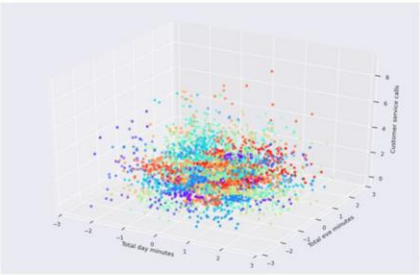
Cluster 2: Represent negative value of day and eve minutes and high customer service calls so they are churner.

DBScan: In DBScan, there were 8 clusters formed. It means it is not performing well on this dataset because it is not dividing data in groups properly.



Cluster	Total day minutes	Total eve minutes	Customer service calls
-1	-0.0	0.1	3.9
0	-0.0	0.0	1.0
1	0.0	0.0	0.0
2	-0.1	-0.0	2.0
3	0.1	-0.1	3.0
4	-0.2	-0.1	5.0
5	0.1	0.1	4.0
6	-0.9	-0.0	7.0
7	0.2	-1.4	5.0

Affinity Propagation: Propagation: In Affinity propagation, 164 clusters were formed. It is performing from DBScan because it is making 164 clusters and not dividing the dataset well.



Cluster	Total day minutes	Total eve minutes	Customer service calls
0	0.6	-0.0	3.0
1	-0.5	-1.6	2.0
2	0.1	-1.5	1.0
3	0.5	-1.0	2.0
4	1.5	-0.9	5.0
...	...	...	...
160	1.0	1.2	1.0
161	-0.4	0.0	2.0
162	1.0	0.5	1.0
163	2.1	0.7	2.0
164	1.1	1.0	3.0

165 rows × 3 columns

RESULTS OF ALL CLUSTERING ALGORITHMS				
Algorithm	**No. of clusters**	**Silhouette Index**	**Calinski Index**	**DB Index**
K-Means	2	0.28	1441.95	1.38
DBScan	8	0.12	301.89	6.24
Affinity Propagation	164	0.32	727.05	0.82
Mean Shift	3	0.43	248.59	1.12

From the above table we analyze that K-mean and Mean shift perform best.

Supervised Learning:

We apply different supervised learning technique using scikit learn library for predictive analysis, from above multivariate analysis we conclude that our data did not contain any pattern and was not linear, so we generate our hypothesis that linear model cannot perform well, Tree based algorithm will perform well in this type of data, we apply Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest also we apply ensemble method bagging algorithms like Adaboost, Gradient Boost, XG Boost, we also apply voting Classifier hard and soft both algorithm in which we input our top algorithm, we also applied Neural Network using Keras tensorflow library but our neural network did not perform well overkill.

We first applied all the algorithms directly without using any imbalance technique and found that our all models were biased; they only predict zero means non churner because our data set contains more samples of non-churner and churner.

So, after analyzing this biased problem, we apply the Imbalance Smote technique to resample our data and then we apply our all above algorithm, after applying Smote all our algorithms perform well to predict both class churner and non-churner except Neural Network.

Model Building:

We first Encode our all categorical variable using label encoder, then separate the Features that are selected from Random Forest from target variable and save into x variable then save target into y variable, random Forest selected 12 features out of 15 features.

Second we apply smote technique to balance our datasets we used sampling strategy is equal to 0.4 and resampling strategy is equal to 0.5,the samples of zero and one before smote is 0: 2850, 1: 483and after applying smote is 0:2280, 1: 1140.

Then we apply train test split of ratio 80/20 and use stratify technique to divide the class into train and test data sets. The size of train and test data set are (2736, 12) (2736) for train and (684, 12) (684,) for test.

After splitting Datasets into train and test we scaled our dataset using sklearn library function Standard Scalar it uses z score transformation technique to scale the data, first we fit and transform on train data set and then we transform into test dataset to avoid data leakage.

After all this preprocessing we applied Various Machine learning Algorithms. The results of this entire algorithm are listed below in the Experiment and Result Section.

At first we applied all algorithms using default parameter ,then we did hyper parameter tuning of all our algorithms in order to optimize our results, we tune our all machine learning models using Randomized Search CV, after tuning parameter we get best parameter for our model ,we apply tune parameter into our models and then analyze the results.

After Hyper parameter tuning we analyze that some of our model are given good results on default parameter and some are after tuning. The Results for all algorithms are listed below in the experimental and results part.

Experiments and Results:

We did Experiments on making two data frames one with all the features and one with the feature selected by Random Forest we analyze that with all features our models did not classify zero and 1 efficiently and there AUC score is also very low, there Miss classification is high. Whereas the feature which is selected by the wrapper method gives us a good AUC score and less misclassification.

Results before applying SMOTE:

Algorithm Name	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score	AUC Score	Misclassification
Logistic Regression	86%	85%	66%	57%	58%	56%	100
Neural Network	95%	87%	43%	50%	46%	50%	97
Support Vector Machine	85%	85%	43%	50%	46%	50%	97
Decision Tree	96%	94%	91%	83%	86%	83%	43
Random Forest	98%	92%	94%	75%	81%	73%	53
AdaBoost	98%	94%	89%	85%	87%	85%	42
Gradient Boost	96%	93%	93%	79%	84%	78%	45

From above results we analyze that our data set is biased so we need to apply some imbalance technique then we applied Smote to resample our data, our data set size was also change 684 test labels

Result after applying SMOTE:

Algorithm Name	Training Accuracy	Test Accuracy	Precision	Recall	F1-Score	AUC Score	Misclassification
Naïve Bayes	77.49%	79%	78%	72%	73%	71%	147
Logistic Regression	76.2%	76%	74%	69%	70%	68%	164
Neural Network	95%	87%	33%	50%	40%	50%	228
Support Vector Machine	73.4%	73.2%	73%	63%	63%	62%	183
Decision Tree	92%	90%	89%	87%	88%	87%	70
Random Forest	97.15%	92.25%	92%	90%	91%	90.01%	53
AdaBoost	100%	88.16%	87%	86%	87%	86%	81

Gradient Boost	94.1%	92%	92%	89%	89%	89%	57
XG Boost	94%	92%	92%	89%	89%	89%	57

From above table we analyze that Gradient Boost and XGBoost perform well, after applying this algorithm we also applied voting algorithm for which we input our top algorithm as a base estimator, the result of voting algorithm is here in below table.

Algorithm Name	Training Accuracy	Test Accuracy	Precision	Recall	F1-Score	AUC Score	Misclassification
Voting Classifier	96%	92%	92%	89%	91%	89.3%	55

After applying voting classifier, we analyzed that XGBoost is performing well to predict churn and non-churner.

Confusion Matrix of XGBoost:

Non-Churner	**True Positive** 439	**False Negative** 17
Churner	**True Negative** 40	**False Positive** 188

Conclusion and Future Work:

In this project, we researched on customer churn prediction by working on telecom dataset. The results were accurate enough to ensure long-term satisfaction and loyalty of the customer with the company. Identifying churners is very important for a company because in terms of costs and advertisement, it's easier to maintain present customers rather than attract new customers. When companies identify the churner company will easily stop the customer to churn by providing him better offers.

The results were accurate but were limited by the lack of the following features, If we were provided with daily transactional data, we would have been able to use RFM analysis (Recency, Frequency, and Monetary value) to produce even more accurate results, profession is also an important factor in order to increase the accuracy it would allow us to understand who would take what offers and would be either a long-term or short-term customer for example a marketing person will have more use of cell phone as compare to Doctor, so if we have the above mentioned features a telecom company can easily stop the churner by offering him attracting packages.

We currently have very good results on the customer dataset which is a good thing for the telecom industry and for making deployment easy we will implement Docker because the great advantage of Docker is that it makes your deployment easy and when the time comes you can even deploy it on multiple machines with almost no extra effort [5]. Secondly we will build an

API so that one can take advantage of our research which we have done in our project and can directly use for more research in the Churn prediction.

References:

- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626-4636.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(3), 90-95.
- Bisong, E. (2019). Matplotlib and Seaborn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 151-165). Apress, Berkeley, CA.
- Piatetsky-Shapiro, G. (2008). KDnuggets.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Mason, L., Baxter, J., Bartlett, P. L., & Frean, M. R. (2000). Boosting algorithms as gradient descent. In *Advances in neural information processing systems* (pp. 512-518).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.
- Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. *Information fusion*, *6*(1), 63-81.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.086*

Group Members:

Team Lead
Name: Maad Saifuddin
Education: Student of Computer Science at DHA Suffa University.
Certification: Data Science Certification from FTI, IBM Data Science Certification from Coursera.
Experience: Working as a research intern on Tweet Topic Detection.
About: I am student of BS(CS) of 7th semester and want to build my career in data field, so I choose data science as my career.

Name: Hussain Murtaza Ali
Education: Student of Computer Science at Sir Syed University of Engineering & Technology.
Certification: Data Science Certification from FTI, data science course from Udacity and Udemy.
Experience: Web Developmen Intern.
About: I am very passionate about data science and want to learn more about data science to build skills and want to pursue data science as my career field.

Name: Muhammad Amir mujahid
Education: BsCs from Iqra University
Certification: Data Science Certification from FTI Data Science, Machine Learning, Data Analysis, Python & R from Udemy.
Experience: Web Developer Intern
About: As a computer science graduate I developed my interest in the field of AI and want to collaborate with professional data scientists in this field to work for my country.

Name: Danial Khan
Education: Computer Engineering from Bahria University Karachi Campus
Certification: Data Science Certification from FTI , Python
Experience: Networking Intern in Pakistan Civil Aviation
About: As a Computer Engineer I studied Artificial Intelligence as a subject then I am curious to learn more because it's an emerging field for next decades and I wanted to pursue my career as a Data Scientist now and make my future in the field of Artificial Intelligence(AI).

Name: Syed Ahmed Ali Naqvi
Education: BSCS from DHA Suffa University
Certification: Data Science Certification from FTI
Machine Learning from Udemy
Android Development from ANA Academy
Experience: Working as a research intern on Tweet credibility Project.
About: As a computer science student I want to pursue my career in Data science field to give the solution of daily life problems faced by society.